Input files
(excel files, .pdf, .csv, etc)

Preprocessing

Data lake (.csv files,
or a Relational DB)

PatientID	Status	Age
PatientID	Status	Age
PatientID	Alive	birthdate

submitter_id	status	age

Target
schema/data

GDC API

case_id	status	age

Domain information

| #____ | Alive | 0-120 |
| #____ | Dead | 0-120 |

Enriched Target
schema/data

case_id		PatientID
#102		101
#101		102

Column Mapping
Domain Mapping

age		birthdate
24		01/01/1990
34		01/01/2000

Column Mapping
Domain Mapping

...		...
...		...
...		...

Column Mapping
Domain Mapping

**Table
Matching**

** Different systems/techniques to implement
this component

PatientID	Status	Age
PatientID	Status	Age
case_id		age

case_id	status	age

**Integration
candidates**

**Table
alignment
and
integration**

** Different systems/techniques to implement
this component

case_id	status	age

**Post
Processing**

case_id	status	age

**Consolidated
data**